DOCUMENT RESUME

ED 151 377 TM 006 795

AUTHOR Filby, Nikola N.
TITLE Progress Report on Reactivity Analyses (October-December Test Data). Beginning Teacher Evaluation Study. Technical Note Series. Technical Note III-5.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
REPORT NO BTES-TN-3-5
PUB DATE 30 Apr 76
NOTE 13p.; For related documents, see TM 006 793 and 794

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Gains; *Achievement Tests; Data Analysis; Data Collection; Effective Teaching; Evaluation Methods; Grade 2; *Item Analysis; Measurement Goals; Primary Education; Reading Tests; Statistical Analysis; Student Evaluation; Teacher Evaluation; *Test Construction; Test Items; *Test Validity; Time Factors (Learning)

ABSTRACT

The development and refinement of the measures of student achievement in reading and mathematics for the Beginning Teacher Evaluation Study are described. The concept of reactivity to instruction is introduced: the tests used to evaluate instructional processes must be sensitive indicators of classroom learning overtime. Data collection activities in second and fifth grade reading and mathematics classes are described. The analysis and evaluation of the reactivity of test items to instruction are summarized. General student ability and the amount of time devoted to instruction are also considered. Several reading test items are included to illustrate the methodology used. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151377

TM006 795

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

BTES

BEGINNING TEACHER EVALUATION STUDY

Technical Note Series

Technical Note III-5

PROGRESS REPORT
ON REACTIVITY ANALYSES
(OCTOBER-DECEMBER TEST DATA)

April 30, 1976

FAR WEST LABORATORY
FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
1855 FOLSOM STREET • SAN FRANCISCO CALIFORNIA 94103 • (415) 565-3000

Technical Note III-5

Progress Report on Reactivity Analyses (October-December Test Data)

by

Nikola N. Filby

April 30, 1976

Overview

A major activity during the Phase III-A Continuation of the Beginning Teacher Evaluation Study has been the further development and refinement of student achievement measures in reading and mathematics. During the winter and spring of 1975 a set of achievement measures was assembled and pilot tested. These initial developmental procedures were reported in BTES Technical Report III-1. (Filby and Dishaw, 1975). This report introduced a central concept in the test development effort -- reactivity to instruction. The tests used to evaluate instructional processes must be sensitive indicators of classroom learning. In order to ensure that the BTES achievement tests will be sensitive, reactive measures of student learning, reactivity data have been collected during the school year 1975-76. This Technical Note is a progress report on the test refinement effort. It summarized the procedures used to collect relevant data and outlines the analysis procedures that are being used to evaluate the reactivity of test items and select the best possible items for future use. A more complete report on refinement procedures and results will be presented in June, 1976.

Data Collection

In order to assess the reactivity of potential test items, tests were administered three times during the 1975-76 school year -- in early October, in December, and in April. Four separate groups of classes were tested -- grade 2 reading, grade 2 mathematics, grade 5 reading and grade 5 mathematics.

In each class, the teacher recorded the content of instruction for individual students for the testing period. In the October - December period, a daily lesson plan log was used. In the December - April period, a weekly content checklist was used. In both cases the teacher records provide information about the amount of time spent working on particular

content areas in the interval between tests. The amount of instructional time received by a student is referred to as "opportunity to learn".

These two sets of data -- test scores and time allocations -- are matched together through a logical analysis of the knowledge and skills assessed by the test items and covered during instruction. That is, teachers reported the content of instruction in categories matching the content of test items. In this way, we can assess the amount of instruction received by a student in categories that are logically relevant to test performance.

Reactivity Analyses

Through the time allocation reports we know what skills a student has had an opportunity to learn and therefore, presumably, those areas in which learning has taken place. Measures of student achievement should be sensitive indicators of this classroom learning. The central question for reactivity analysis is "Do items show gain when instruction has taken place?" Students should perform better on an item after instruction than they do before instruction. Moreover, students who receive a large amount of instruction in a particular area should tend to perform better, show greater gains, than students who receive only a little instruction in a particular area. Items which conform to this pattern are sensitive, reactive measures of classroom learning.

In contrast to this pattern, test items may require or reflect high amounts of general ability. Aspects of test format or test-taking conditions may favor students with higher general ability. Or high ability students may be better able to infer the purpose of test items and apply their relevant knowledge and skill to those items. One kind of evidence that such factors are operating would be a large difference in performance between high and low ability students at the initial October testing session.

Similarly, if there are large differences between ability levels when students have had little opportunity to learn, this suggests that factors other than or in addition to, classroom learning are strongly influencing performance.

Given the concerns outlined above, the general plan for the reactivity analyses is to examine several characteristics of test performance over time. We will select for inclusion in the final test battery, subtests and items which maximize reactivity to the amount of instruction and minimize the influence of general ability factors.

A Procedural Example - Long Vowels

These reactivity analyses can be illustrated with the data for one subtest in grade 2 reading -- long vowels. The data presented will be for the October - December interval. The test items themselves are shown in Appendix A.

In an introduction to the data, it is useful to ask first how much gain occurred from October to December and what distributions of time allocation occurred. On the 22 item long vowel subtest, the difference between the mean score in October and the mean score in December was 3.4 items, for an average gain of 15%.

Time allocations varied widely across students and classes. The range of allocated time was from 0 to 400 minutes, with an overall average of 133 minutes.

The basic question for the reactivity analysis is the extent to which variation in test score gain is associated with differences in opportunity to learn. Does gain occur primarily for students with substantial exposure to instruction and thereby reflect classroom learning. In order to examine

this issue, data are displayed in an Ability-by-Opportunity-to-Learn Matrix, as shown in Table 1.[1]

T73STRAT

MEAN / COUNT / STD DEV	LE 10 MIN	GT 10,LE 60 MIN	GT 60,LE 120 MIN	GT 120,LE 240 MIN	GT 240 MIN	ROW TOTAL
	1	2	3	4	5	
ACSTAT2						
2	5,1000	3,0625	3,3750	3,5500	5,7000	3,8810
GT 26,LE 80	5	8	4	20	5	42
	3,9115	5,0103	3,3250	4,8906	1,1511	4,3122
3	3,0000	1,0833	3,6000	3,1923	2,5714	2,6378
GT 80,LE 160,	1	6	10	13	7	37
	0,0	4,4319	4,0743	2,7122	2,6992	3,3584
COLUMN TOTAL	4,7500	2,2143	3,5357	3,4091	3,8750	3,3924
	6	14	14	33	12	79
	3,6021	4,7015	3,7492	4,1221	2,6554	3,9057

Table 1

Ability - Opportunity-to-Learn Matrix
For Long Vowel Subtest

The rows in this table represent students of two different levels of performance on a measure of general academic status entering grade 2. Students in the top row show lower overall performance in reading at the beginning of grade 2 than students in the bottom row.

The columns in the matrix represent different amounts of time spent working on long vowel decoding. The times are in minutes, with column 1 representing less than 10 minutes of instruction in long vowels and column 5 representing more than 240 minutes of instruction in long vowels.

The values within each cell, reading from top to bottom, indicate the mean gain, the number of students, and the standard deviation of the gain

[1] All data reported here is for a restricted range of overall entering achievement level. In keeping with the Proposal for Phase III-B, students at the top and bottom of the distribution have been eliminated from analyses in order to focus attention on students like those to be included in Phase III-B. This will allow us to select measures which are particularly appropriate and reactive for that group.

scores for students in that cell.

Interpreting the pattern of gain scores in this table, it is most readily apparent that students of low entering status learn more between October and December than students of higher entering status. In part, this is the result of an emerging ceiling effect. At the December testing, 5 students in the higher strata answered all long vowel items correctly and 7 students missed only 1 item. But even without reaching the literal ceiling of the test, students showed a slower rate of growth in the upper end of the distribution. Throughout the analyses, we will be alert to the possibility of ceiling effects. We will try to select or extend subtests to minimize such restrictions. But it may still be the case that growth in phonics skills is greatest for students entering at lower levels. In light of our concern about the opposite case, gain only for high aptitude students, this appears to be an acceptable, even encouraging, pattern.

The pattern of relationship between gain and time is somewhat disconcerting. An inexplicably high gain was shown by students with virtually no direct instruction in long vowels (column 1). Supplementary analyses are underway to try to understand this phenomenon. It appears not to be related to entering status. It is more encouraging to note that the lowest gains are shown by students with relatively little time in long vowel instruction. This dip in the curve provides some evidence of reactivity to instruction. A real evaluation of the magnitude of the reactivity effect will depend on a comparison with other subtest profiles and with December-April patterns.

An analysis of the data at a subtest level provides a useful summary of trends. It may be even more important, through, to evaluate the reactivity of individual items within a subtest. In a follow-up to the

subtest matrix in Table 1, individual items were displayed in an abbreviated Ability-OTL matrix. Sample item displays are shown in Table 2.

	T73STRAT			
MEAN COUNT STD DEV	LE 10 MI N	GT 10,LE 60 MIN	GT 60 MIN	ROW TOTAL
	1	2	3	
ACSTAT2				
2	0,0	0,13	0,17	0,14
GT 26,LE 80	5	8	29	42
	0,0	0,35	0,38	0,35
3	0,0	0,0	0,07	0,05
GT 80,LE 160,	1	6	30	37
	0,0	0,0	0,52	0,47
COLUMN TOTAL	0,0	0,07	0,12	0,10
	6	14	59	79
	0,0	0,27	0,46	0,41

Positive Pattern Item

	T73STRAT			
MEAN COUNT STD DEV	LE 10 MI N	GT 10,LE 60 MIN	GT 60, MIN	ROW TOTAL
	1	2	3	
ACSTAT2				
2	0,40	0,13	0,10	0,14
GT 26,LE 80	5	8	29	42
	0,55	0,35	0,31	0,35
3	0,0	0,17	0,0	0,03
GT 80,LE 160,	1	6	30	37
	0,0	0,41	0,0	0,16
COLUMN TOTAL	0,33	0,14	0,05	0,09
	6	14	59	79
	0,52	0,36	0,22	0,29

Negative Pattern Item

Table 2

Sample Item Matrices

The layout of the item tables is the same as the subtest table. The only difference is that the three largest time categories have been collapsed into a single category.

The top table shows the pattern of gain with opportunity that one would expect to see of a reactive item. This is labeled as a positive pattern. The bottom table shows the reverse trend and is labeled a negative pattern.

It is worth noting that these opposing patterns occur for items with approximately equal overall gain. In each case there is the number of students responding correctly.

Reactivity data for all 22 items in the long vowel subtest are summarized in Table 3. This table indicates the proportion of students responding correctly in October and the percent gain from October to December. In addition, the pattern of the relationship between gain and time is indicated. Some items showed patterns that were clearly positive or negative, as shown in the sample tables. Other items showed unclear or mixed (M) patterns.

On the basis of these three kinds of information, items have been divided into three groups. First choice items are those with large gains clearly related to opportunity to learn. These items will be used in the Phase III-B battery. Second choice items show less strong but desirable characteristics. These items will be used if it is decided to retain a full subtest of long vowel items in the final battery. The remaining items will probably be deleted or replaced.

It is interesting to compare the reactivity properties of individual items with standard item statistics. The last column in Table 3 indicates the point-biserial correlation of each item with the subtest total score.

Table 3

Reactivity Data
for Items in the Long Vowel Subtests
(Grade 2 Reading)

Status	Item Number	Proportion of Students Responding Correctly in October	Proportion Gain October to December	Pattern of Gain with OTL	Point-Biseral Correlation with Subtest Scores (Dec.)
First	1	.81	.10	+	.30
	4	.54	.24	+	.59
Choice	14	.60	.20	+	.81
	15	.40	.23	+	.67
	17	.47	.20	M	.67
	18	.52	.10	+	.73
Second	3	.61	.16	M	.55
	8	.78	.10	M	.69
Choice	11	.39	.04	+	.59
	13	.51	.10	M	.66
	19	.88	.06	+	.32
	20	.51	.13	+(c)	.55
	21	.67	.14	M	.72
Probable	2	.86	.09	-	.40
	5	.82	.06	-	.38
Rejects	6	.57	.06	-	.43
	7	.86	.01	-	.59
	9	.77	.03	M	.67
	10	.50	.10	-	.55
	12	.57	.03	M	.66
	16	.89	.08	-	.21
	22	.67	.01	M	.67

Standard test development procedures would favor items with a high correlation. These items are not necessarily the most reactive.

Future Analyses

The procedures described above will be carried out for all subtests and items in the trial test battery. Results for the October - December period will be compared with results from the December - April period. Some attention will be paid to a comparison of between-class and within-class analyses. Also, analyses will be carried out at different levels of specificity in content definition (long vowels versus total decoding) in order to identify the level of specificity necessary for future teacher record keeping.

On the basis of reactivity data and logical analysis, subtests and items will be selected for use in the Phase III-B battery. Selected items, along with the rationale for their selection, will be presented in Technical Report III-6, scheduled for completion June 30, 1976.

Appendix A

Long Vowel Items

Circle the word that names the picture.

33	beds beads baids	38	soup soap sop
34	goot got goat	39	heat hat hait
35	peal pal pail	40	chain chean chan
36	ten teen tien	41	fet foet feet
37	lef loaf leaf	42	coat cot coot

GO ON TO THE NEXT PAGE